J. Luo

CSC 240/440 Data Mining
NO SMARTPHONES - YOU DON'T REALLY NEED A CALCULATOR

Name: Xuejian Ye **Department:** Data Science **240/440** 440

CSC240/440 Data Mining: 2017 Midterm Exam Questions (10 pts each)

1. (a) In a study published on 10/30/2017 in *Nature Human Behavior*, researchers at Carnegie Mellon and the University of Pittsburgh analyzed how suicidal individuals think and feel differently about life and death, by looking at patterns of how their brains light up in an fMRI machine. With 17 subjects, an accuracy of 87% is reported. How significant is this finding? And why? (b) Recently, AlphaGo Zero crushed AlphaGo which defeated human champions by learning to play the game without using any data from human plays. It was touted as an example of self-learning by machines without data. Is this true? And Why?

(a) this finding is ~~not very~~ meaningful, but suicidal individuals is a very small part in human. So the accuracy of 87% is not sufficient enough. Besides, the subject sample is small, So this finding still need some further research.

(b) NO, It's an unsupervised method to train this machines, but It still need some ~~data, because~~ they need to know the rules of playing the game.

2. Suppose a group of 12 students with the test scores listed as follows:
26, 51, 48, 63, 35, 85, 69, 81, 72, 88, 44, 95. 26, 35, 44, 48, 51, 63, 69, 72, 81, 85, 88, 95
(a) Partition them into four bins by (1) equal-frequency (equi-depth) method, (2) equal-width method, and (3) an even better method (such as clustering). Which is better and why?

(1) bin1 : (26, 35, 44)
bin2 : (48, 51, 63)
bin3 : (69, 72, 81)
bin4 : (85, 88, 95)

(2) width = $\frac{95-26}{4} = \frac{69}{4} = 17.25$
bin1 : (26, 35, 44)
bin2 : (48, 51, 63)
bin3 : (69, 72, 81)
bin4 : 85, 88, 95)

(3) I prefer to use k-means method because It's ease to implement and has a good cluster efficiency.

(b) What are the value ranges of the following normalization methods, respectively?
(1) min-max normalization, (2) z-score normalization, and (3) normalization by decimal scaling?

(1) min-max normalization is for any ranges of value.
$\frac{V_i - 26}{95-26} * [0, 1]$

(2) range of z-score normalization : $\left[\frac{min_A - \bar{A}}{6}, \frac{max_A - \bar{A}}{6}\right]$, $\bar{A} = \frac{1}{n}[V_1 + V_2 + \cdots + V_i$

(3) range of normlization by decimal scaling : $\left[\frac{min}{10^j}, \frac{max}{10^j}\right]$, $(max = \frac{V_i}{10^j}) < 1$
$\therefore \left[\frac{26}{10^2}, \frac{95}{10^2}\right]$

use data.

CSC 240/440 Data Mining J. Luo

NO SMARTPHONES - YOU DON'T REALLY NEED A CALCULATOR

+10

3. Suppose you have the following information:
Today is Halloween. 70% of people on UR campus are students, 15% are faculties. Last year, 30% of students, 10% of faculties, and 20% of other UR staff wore costumes on Halloween, respectively.
(1) What is the probability of seeing a person in costume on the UR campus today?
(2) What is the a-posteriori probability that a person is a faculty given that s/he is not in costume today?

(1) $P(C) = P(stu) \cdot P(C_1|stu) + P(fac) \cdot P(C_2|fac) + P(other) P(C_3|other)$
$= 0.7 \times 0.3 + 0.15 \times 0.1 + 0.15 \times 0.2$
$= 0.255$

(2) $P(fac|C^c) = P(fac) \cdot \frac{P(C^c|fac)}{P(C^c)}$, $[P(C^c) = 1 - P(C) = 1 - 0.255 = 0.745$, $P(C^c|fac) = 1 - P(C|fac) = 1 - 0.1 = 0.9]$

$\therefore P(fac|C^c) = 0.15 \times \frac{0.9}{0.745} = \frac{27}{149}$

6

4. Basics of data mining.
(a) What are the best distance measure for each of the following applications?
(i) Delivering express mails in Downtown LA

Manhatan measure. distance $= \sum |X_{ia} - X_{jb}|$

(ii) Finding similar news in Twitter and New York Times

Minkinsic measure distance $= \sqrt[h]{\sum_{i,j}^{m} (X_{ia} - X_{jb})^h}$

(iii) Calculating the fuel cost of transatlantic flights

Euclide measure (欧几里德) $dis = \sqrt{(X_{i1} - X_{j1})^2 + (X_{i2} - X_{j2})^2 + \cdots (X_{ia} - X_{jb}}$

(iv) Comparing diabetes with a set of medical tests

$Max[X_{ia} - X_{jb}]$

(b) Explain the following terms frequently used in data mining, what it means / what it does?

PCA Principal component analysis
ID3 a decision-tree algorithms
ROC a picture contain the curve of TPR and FPR, which is used to estimate the result of clas
DBSCAN a density-based cluster ~~antigram method~~ algorithms

(c) Name and describe one method that perform effective dimensionality reduction **and** one method that perform effective numerosity reduction.

numersity reduction is deleted the null value and the noise, outlier.

dimensionality Reduction: use close fitemset and maximum itemset.

5. Given a fixed min support threshold, σ (e.g., σ = 0.5%), present an efficient incremental Apriori algorithm that can *use the previously mined information* without re-examining the early data TDB when a new set of transactions ΔTDB is added.

7

I will use Apriori algorithm with partitianing, the dataset TDB and new set ΔTDB, we can view as two partitions, then find the local frequent itemset to each partitions and each partition has their own minimum support. Then combine the two part frequent itemset into one ~~comb~~ candidate itemset and reuse Apriori algorithm to find the global frequent itemset, which is the final frequent itemset of the combination of TDB and ΔTDB.

6

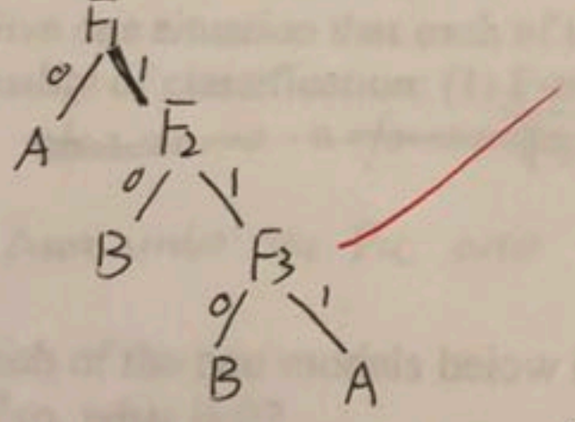
6. Given the following training examples.

F1	F2	F3	CLASS
1	1	1	A
1	1	0	B
0	1	0	A
1	0	1	B

6

(a) Work out one decision tree that is deeper than three levels. What criterion does this correspond to?

(a)



This correspond to the info-gain criterion, $info(D) = -\sum_{i=1}^{m} P_i \cdot \log_2 P_i$

$info(D_1) = \left|\frac{P_1}{D}\right| \sum_{i=1}^{m} P_i \log_2 P_i$, $Gain = info(D) - info(D_i)$

if Gain is biggest, then this attribution is choosed.

0

(b) How would you avoid overfitting when constructing a decision tree for big data? Briefly describe the process.

we can do better in the process of data clean.
minimize the influence of noise and outlier.

CSC 240/440 Data Mining J. Luo

NO SMARTPHONES - YOU DON'T REALLY NEED A CALCULATOR

7. Understanding classification performance.

(a)What are the major differences among the following two methods for the evaluation of the accuracy of a classifier: (1) hold-out method, (2) stratified cross-validation?

(1) hold-out method is divide the dataset into two part, $\frac{2}{3}$ of dataset as train set, $\frac{1}{3}$ of dataset as test set, all the data is choosed only one time.

(2) ~~or~~ stratified cross-validation is divide the dataset into K part with equal size. $(D_1, D_2, \cdots D_K)$ Interative, when D_1 is test set, the remains of data become train set. All the data will be used many times.

(b) Calculate the performance metrics from the following confusion matrix:

		Truth A	Truth B	Truth C	
Classification	A	90	9	1	100
	B	6	86	8	100
	C	4	5	91	100
		100	100	100	300

+9

Sensitivity (true positive rate) for each class

$$\text{Sensitivity} = \frac{TP}{TP+FN} \qquad TPR_{(A)} = \frac{90}{100} = 0.9 \;, \; TPR_{(B)} = \frac{86}{100} = 0.86, \; TPR_{(C)} = \frac{91}{100} = 0.91$$

Specificity (true negative rate) for each class

$$\text{specificity} = \frac{FP}{FP+TN} \qquad Spec_{(A)} = \frac{90}{200} \;, \; Spec_{(B)} = \frac{86}{200} \;, \; Spec_{(C)} = \frac{91}{200}$$

Recall for each class

$$\text{Recall} = \frac{TP}{TP+FN} \qquad R_{(A)} = 0.9 \;, \; R_{(B)} = 0.86, \; R_{(C)} = 0.91$$

Precision for each class

$$\text{Precision} = \frac{TP}{TP+FP} \qquad Pre_{(A)} = 0.9 \quad Pre_{(B)} = 0.86 \quad Pre_{(C)} = 0.91$$

8. Evaluation of Classification.

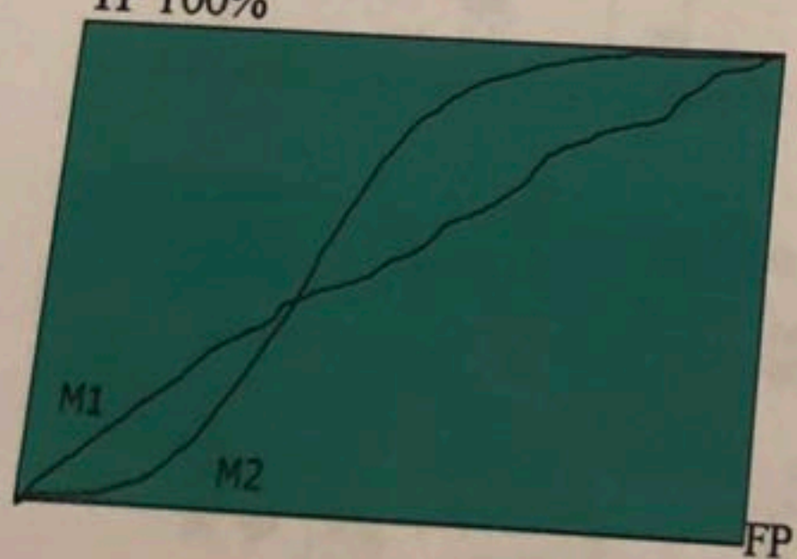
(a) Give one situation that each of the following measures is most appropriate for measuring the quality of classification: (1) F-measure, and (2) Area under the ROC curve.

4

(1) ~~when we use a famous Parties~~ whether a news about a sport star is a sport news because some News about a sport star is not a sport News we need to calculate the TP, FP, TN, FN which is used by F-measure and Roc curve

(2) Area under the Roc cure means the accuracy of this classifies.

(b) Which of the two models below is better? Is there a reason not to use the model you just pick? If so, what is it?



M_2 is better, compare with M_1, the area under the Roc curve is bigger, which means the accuracy of M_2 is better, and the line of M_1 is too close to the diagnals, which means the accuracy of M_1 is not good enough.

CSC 240/440 Data Mining J. Luo

NO SMARTPHONES - YOU DON'T REALLY NEED A CALCULATOR

9. What is class-imbalance problem?
(a) Name at least FOUR different strategies for alleviating the problem
(b) What strategy could you device to utilize ALL the data samples you have in a class-imbalance problem?

(4)

(a) undersampling, oversampling, ensemble method, threshold-moving

(b) I prefer to use ensemble method, which is combines both function of undersampling and ~~threshold-or~~ threshold-moving. Its a good method to deal with class-imbalance whether the size of data samples.

9

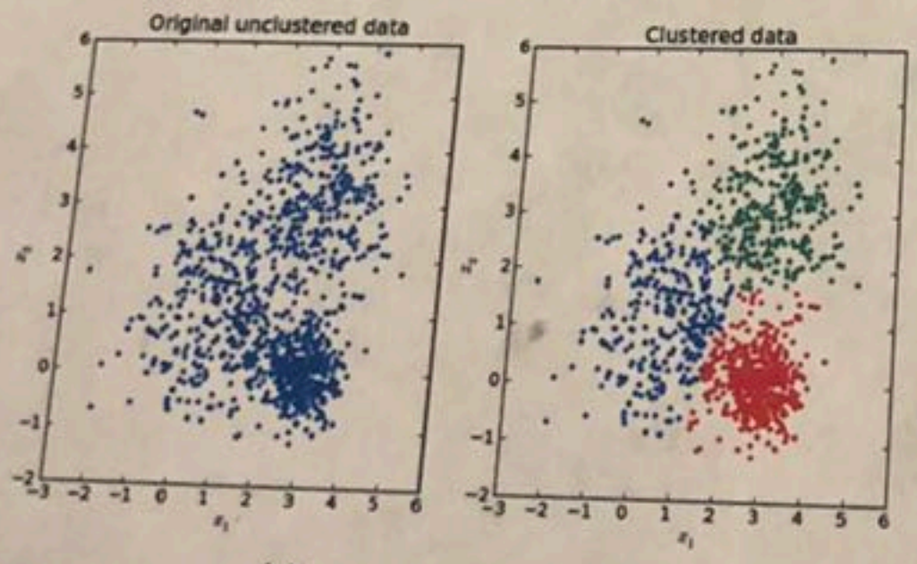
10. Clustering Analysis.
(1) What are the four primary approaches to clustering? Name an example algorithm using each of these approaches.

(7)

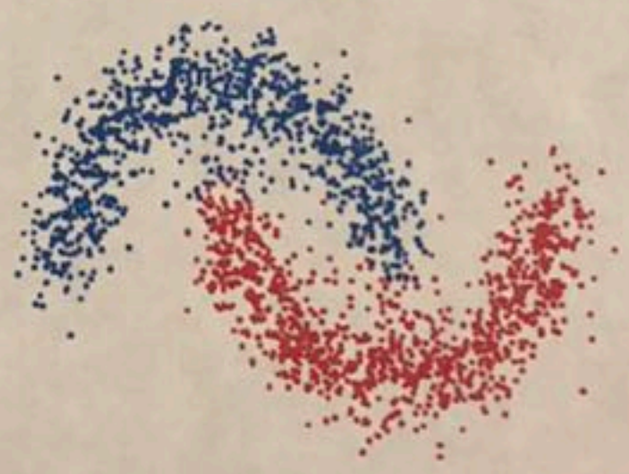
(A) partition (k-means)
(B) ~~hierarchy~~ hierarchy (Birch)
(C) density-based method (DBSCAN)
(D) grid-based method (STING)

4

(2) Different data sets may require (i) different similarity (distance) measures, along with (ii) different clustering algorithms. Propose a 'good' solution for each of the following datasets (a) and (b) to achieve the desired clustering results, respectively. Specify both the **distance measures** and the clustering **algorithms**, and explain **why** your solutions are 'good'.



(a)



(b)

(a) distance measure: ~~distance~~ $distance_{min} = Min[V_i - P_i]$
algorithms: partition (k-~~means~~ medoid)
because ~~distance~~ Min is a efficient way for us to find the closest core P_i, which is crucial part for k-medoid.

(b) distance measures: ~~Euri~~ Average distance.
algorithms: density-based (DBSCAN)
because density-based could find a cluster with arbitrary shape.

1. Circle which of the following are data mining tasks? (3

a) Sorting a student database based on student identification numbers.

b) Monitoring the heart rate of a patient for abnormalities

c) Computing the total sales of a company

d) Store all data in an Excel file

e) Extracting the frequencies of a sound wave

f) Monitoring seismic waves for earthquake activities.

4'

2. Classify the following attributes as binary, discrete, or con
Also classify them as nominal, ordinal, interval, or ratio.

Example: Age in years. Answer: discrete, ratio.

a) Bronze, Silver, and Gold medals as awarded at the Olympics.

discrete nominal

b) Number of patients in a hospital.

discrete. nominal

c) Military rank.

discrete ordinal

d) Brightness as measured by a light

4. Naïve Bayes Classifier

(10 points)

Consider the following data set with Attributes A, B, C and class label "-" and "+" .

Index	A	B	C	Class
1	0	0	1	-
2	1	0	1	+
3	0	1	0	-
4	1	0	0	-
5	1	0	1	+
6	0	0	1	+
7	1	1	0	-
8	0	0	0	-
9	0	1	0	+
10	1	1	1	+

(a) Predict the class label for a test sample (A = 1, B = 1, C = 1) using the naive Bayes approach

(8 points)

$X = \{A=1, B=1, C=1\}$

$l_i = \{"-", "+"\}$

$P(l_i|X) = \frac{P(X|l_i)\cdot P(l_i)}{P(X)}$ ← To maximum l_i's, according to naive Bayes approach

$P(l_{"-"}) = \frac{5}{10} = \frac{1}{2}$

$P(l_{"+"}) = \frac{5}{10} = \frac{1}{2}$ $\Rightarrow P(l_{"-"}) = P(l_{"+"})$

$P(C=1|l_{"-"}) = \frac{1}{5}$ $P(A=1|l_{"-"}) = \frac{2}{5}$

$P(C=1|l_{"+"}) = \frac{4}{5}$ $P(A=1|l_{"+"}) = \frac{3}{5}$

$P(B=1|l_{"-"}) = \frac{2}{4} = \frac{1}{2}$

$P(B=1|l_{"+"}) = \frac{1}{2}$

$P(l_{"-"}|X) = \frac{2}{5} \times \frac{1}{2} \times \frac{1}{5} = \frac{1}{25}$

$P(l_{"+"}|X) = \frac{3}{5} \times \frac{1}{2} \times \frac{4}{5} = \frac{6}{25}$

$P(l_{"+"}|X) > P(l_{"-"}|X)$

∴ Classify as Class "+" for (A=1, B=1, C=1)

(b) What is an assumption when using the Näïve Bayes classifier? (2 points)

The Conditional Probability $P(X|C_i)$ inside each attribute are independent.

PLEASE KEEP YOUR EYES ON YOUR OWN PAPER